中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

高精度实时语义分割算法框架：多通道深度加权聚合网络

齐咏生, 陈培亮, 高学金, 董朝轶, 魏淑娟

引用本文:
齐咏生,陈培亮,高学金,董朝轶,魏淑娟. 高精度实时语义分割算法框架: 多通道深度加权聚合网络[J]. 控制与决策, 2024, 39(5): 1450−1460.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1699

高精度实时语义分割算法框架: 多通道深度加权聚合网络

齐咏生[1,2,3†], 陈培亮[1,2,3], 高学金[4], 董朝轶[1,2,3], 魏淑娟[1,2,3]

(1. 内蒙古工业大学 电力学院，呼和浩特 010080；2. 大规模储能技术教育部工程研究中心，呼和浩特 010080；3. 内蒙古自治区高等学校智慧能源技术与装备工程研究中心，呼和浩特 010080；4. 北京工业大学 信息学部，北京 100080)

摘 要: 近年来随着深度学习技术的不断发展,涌现出各种基于深度学习的语义分割算法,然而绝大部分分割算法都无法实现推理速度和语义分割精度的兼得. 针对此问题,提出一种多通道深度加权聚合网络(MCDWA_Net)的实时语义分割框架. 该方法首先引入多通道思想,构建一种3通道语义表征模型,3通道结构分别用于提取图像的3类互补语义信息:低级语义通道输出图像中物体的边缘、颜色、结构等局部特征;辅助语义通道提取介于低级语义和高级语义的过渡信息,并实现对高级语义通道的多层反馈;高级语义通道获取图像中上下文逻辑关系及类别语义信息. 之后,设计一种3类语义特征加权聚合模块,用于输出更完整的全局语义描述. 最后,引入一种增强训练机制,实现训练阶段的特征增强,进而改善训练速度. 实验结果表明,所提出方法在复杂场景中进行语义分割不仅有较快的推理速度,且有很高的分割精度,能够实现语义分割速度与精度的均衡.

关键词: 深度学习；语义分割；语义特征；上下文信息；深度融合

中图分类号: TP183 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1699

引用格式: 齐咏生,陈培亮,高学金,等. 高精度实时语义分割算法框架: 多通道深度加权聚合网络[J]. 控制与决策, 2024, 39(5): 1450-1460.

High precision real-time semantic segmentation algorithm: Multi-channel deep weighted aggregation network

QI Yong-sheng[1,2,3†], *CHEN Pei-liang*[1,2,3], *GAO Xue-jin*[4], *DONG Chao-yi*[1,2,3], *WEI Shu-juan*[1,2,3]

(1. School of Electric Power, Inner Mongolia University of Technology, Hohhot 010080, China；2. Engineering Research Center of Large Energy Storage Technology of Ministry of Education, Hohhot 010080, China；3. Center for Intelligent Energy Technology and Equipment Engineering, Inner Mongolia University, Hohhot 010080；4. Faculty of Information Technology, Beijing University of Technology, Beijing 100080, China)

Abstract: In recent years, with the continuous development of deep learning technology, various semantic segmentation algorithms based on deep learning have emerged, but most of the segmentation algorithms cannot achieve high speed and high accuracy at the same time, and a real-time semantic segmentation framework for multi-channel depth-weighted aggregation networks (MCDWA_Net) is proposed to solve this problem. Firstly, the multi-channel idea is introduced to construct a three-channel semantic representation model, which is used to extract three types of complementary semantic information of the image: 1) Low-level semantic channel outputs the local features such as the edge, color, and structure of the object in the image; 2) Auxiliary semantic channel extracts the transition information between low-level semantics and high-level semantics, and realizes multi-layer feedback to the high-level semantic channel; 3) Advanced semantic channel obtains context logical relationships and category semantic information in images. Then, a three-class semantic feature weighted aggregation module is designed to output a more complete global semantic description. Finally, an enhancement training mechanism is introduced to realize the feature enhancement in the training stage, thereby improving the training speed. Experimental results show that the proposed method not only has fast inference speed, but also has high segmentation accuracy in complex scenes, which can achieve the balance of semantic segmentation speed and accuracy.

Keywords: deep learning；semantic segmentation；semantic feature；context information；depth fusion

收稿日期: 2022-09-26；**录用日期:** 2023-02-10.

基金项目: 国家自然科学基金项目(62241309)；内蒙古科技计划项目(2020GG028, 2021GG164)；内蒙古自然科学基金项目(2020MS05029, 2021MS06018).

责任编委: 高会军.

†通讯作者. E-mail: qys@imut.edu.cn.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

0　引　言

近年来,随着深度学习技术的不断发展, 图像语义分割算法已成为一种重要的智能感知技术,其任务是为每个像素分配语义标签,将图像中不同物体的像素区域分隔开,并对每一块区域的类别进行标注[1]. 当前,语义分割技术具有广泛的应用价值,如场景理解[2]、自动驾驶[3]、车辆识别[4]、人机交互和视频监控等.

在基于深度学习的语义分割算法中,一些方法通过限制图像大小[5]或修剪网络的冗余通道来降低网络的计算复杂度[6],以此提高算法推理速度,实现实时语义分割,但这会使此类实时语义分割算法的精度降低;也有一些方法利用U形结构在高分辨率特征图上进行操作[7],提高分割精度,但这降低了算法推理速度. 考虑到单纯基于深度学习的语义分割算法在速度和精度平衡上能力有限,算法性能仍有较大提升空间,为此一些学者提出了从提高语义分割算法的学习能力上来改进思路. 其中以神经进化方法[8]最为显著,如神经进化卷积神经网络[9]、生成式对抗网络[10]、自动编码器网络[11]、长短期记忆网络[12]、深度强化学习[13]等典型模型,这些模型能较好地提升算法的学习能力,在推理速度上仍有缺陷,即无法从本质上解决实时推理速度与算法分割精度的矛盾.

针对此问题提出的改进措施有:文献[14]提出的BiSeNet语义分割算法利用双分支结构结合空间信息在保证一定速度的同时,尽可能提高算法的分割精度,取得了一定的效果. 但分支之间的相互独立性限制了算法的学习能力,使算法无法更好地学习图像的全局语义特征,使得分割精度的提升有限. 文献[15]利用多分支结构提出了DFANet语义分割算法,该算法在一定程度上提高了算法的分割精度. 不过该多分支结构均完全相同,因此本质上只能提取一类语义信息,无法更好地获取全局语义和上下文信息特征,使得该算法受到了一定的局限.

针对上述问题,本文进一步深化多通道思想,并全新层次化定义了语义信息,提出一种多通道深度加权聚合网络框架(multi-channel deep weighted aggregation net, MCDWA_Net),实现高精度实时语义分割. 其主要贡献如下:MCDWA_Net网络框架由3通道层次化语义表征模块、3类语义特征加权聚合模块和增强训练机制模块组成. 其中,3通道层次化语义表征模块用于提取更全面的图像特征,包括低级、辅助、高级3类互补语义信息,并加快图像特征提取速度;3类语义特征加权聚合模块实现互补语义特征加权后深度融合,输出全局语义特征,可大幅提高网络的分割精度;此外,为增强训练机制速度,提出一种新的损失函数,可有效增强训练阶段的特征表达,强化和改善训练速度.

1　理论基础

1.1　轻量结构实时语义分割算法

自全卷积语义分割(fully convolutional networks for semantic segmentation, FCN)算法[16]被提出后,又在此基础上涌现出各种新型的语义分割算法,主要为扩展主干网络[17]和编码器-解码器主干网络[18],但这些网络的实时性几乎都难以满足不断提升的应用需求.

近年来,为提高语义分割算法推理速度提出了多分支网络结构. 如文献[19]提出的BiSeNet V2语义分割网络框架,如图1所示,用两个通道分别定义图像语义信息,然后将其输出结果通过聚合方法融合,最终输出完整的图像语义信息. 该算法推理速度较快,在实际应用中无法对未知环境中的物体进行高精度的分割,具有局限性.

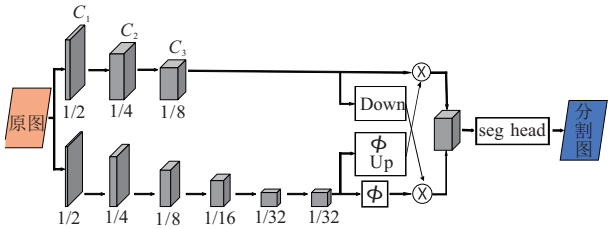


图 1　BiSeNet V2网络框架

1.2　深度可分离卷积与普通卷积

深度可分离卷积(depthwise separable convolution, DSC)[20]是目前轻量化网络中运用最广泛的一种卷积方式,其主要作用是缩减网络参数,从而加快网络推理速度.

深度可分离卷积用于提高算法运行速度主要分为两步:逐通道卷积和逐点卷积,如图2(a)和图2(b)所示,普通卷积示意图如图2(c)所示.

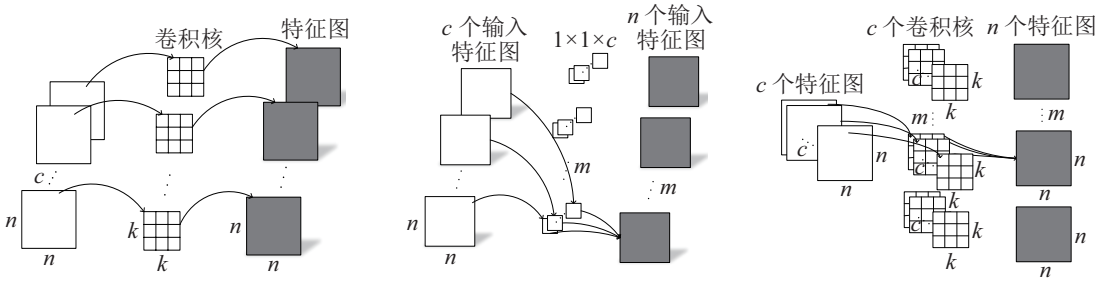
对输入的c个特征图,计算DSC核数量为

$$\mathrm{Conv}_{\mathrm{DSC}} = ck^2 + cm. \tag{1}$$

其中:k为卷积核大小,m为逐点卷积特征图数目.

设输入特征图矩阵为$I_{n\times n}$,层数$C_{\mathrm{in}} = c$,输出特征图层数$C_{\mathrm{out}} = m$,卷积核大小$k = 3$,卷积步长$s = 1$,则卷积过程可表示为$\mathrm{Conv}(I_{n\times n}, C_{\mathrm{in}}, C_{\mathrm{out}}, k, s)$,普通卷积核数量为

$$\mathrm{Conv}_{\mathrm{normal}} = cmk^2. \tag{2}$$



(a) 逐通道卷积 (b) 逐点卷积 (c) 普通卷积

图 2 通道卷积

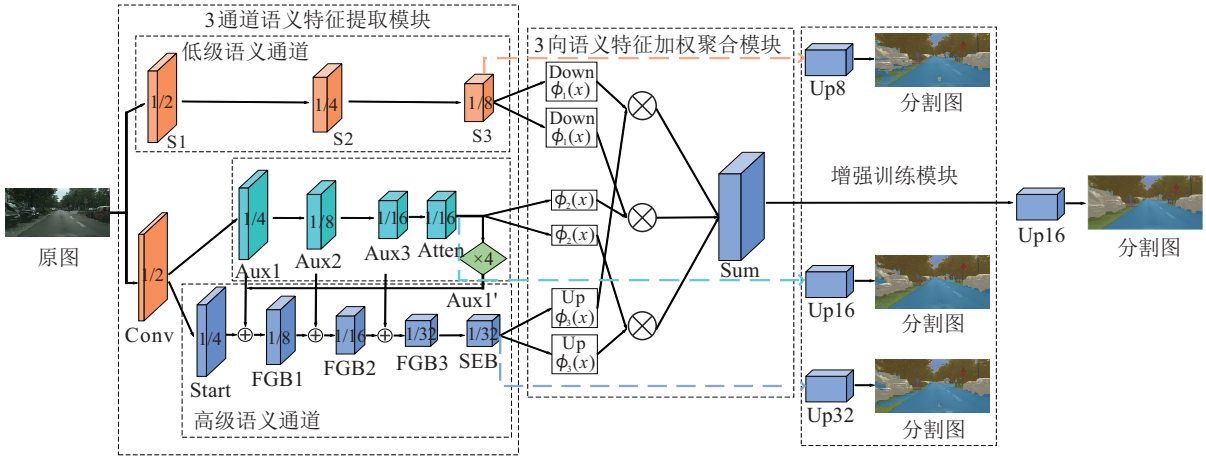


图 3 MCDWA_Net框架

显然$\text{Conv}_{\text{DSC}} < \text{Conv}_{\text{normal}}$,即深度可分离卷积的卷积核数量少,运行速度更快.

2 一种高性能实时语义分割网络框架

2.1 MCDWA_Net框架结构介绍

图3为MCDWA_Net框架总体示意图,3个主要部分为:3通道层次化语义表征模块、3类语义特征加权聚合模块和增强训练机制模块.

2.2 3通道层次化语义表征模块

2.2.1 低级语义通道

低级语义通道主要用来提取图像中物体类别的边缘、颜色、结构等局部特征语义信息,而浅层宽通道网络结构对该类信息比较敏感,为此,设计一种基于浅层宽通道结构卷积网络实现低级语义通道的特征提取,如图4所示. 浅层结构卷积网络的优势是结构简单、网络推理速度快.

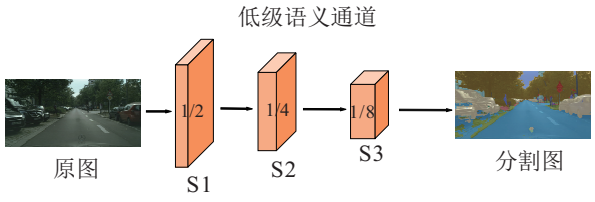


图 4 低级语义通道网络结构

设输入图像矩阵为$I_{H\times W}$,通道数为3,低级语义信息的提取为表1中S1 ~ S3的过程. 表1中,H和W分别为输入图像的高和宽,Conv表示卷积操作,k为卷积核大小,C_{out}为输出图像通道数,s为卷积步长.

设函数$\text{LS}(x)$为低级语义通道信息提取过程,则该低级语义信息提取过程如下所示:

$$\text{LS}(I_{H\times W}) = \text{S3}(\text{S2}(\text{S1}(I_{H\times W}))). \tag{3}$$

由表1知,低级语义通道最终输出$H/8 \times W/8 \times 64$的低级语义特征$\text{LS}(I_{H\times W})$,能更好的表达图像中局部特征的语义信息.

表 1 低级语义通道语义信息提取过程

过程	输入尺寸	低级语义信息提取				
		操作	k	C_{out}	s	输出尺寸
S1	$H\times W$	Conv	3	16	2	$H/2\times W/2$
	$H/2\times W/2$	Conv	3	16	1	$H/2\times W/2$
S2	$H/2\times W/2$	Conv	3	32	2	$H/4\times W/4$
	$H/4\times W/4$	Conv	3	32	1	$H/4\times W/4$
S3	$H/4\times W/4$	Conv	3	64	2	$H/8\times W/8$
	$H/8\times W/8$	Conv	3	64	1	$H/8\times W/8$

2.2.2 辅助语义通道

辅助语义通道主要用来提取图像中介于低级语义和高级语义的过渡语义信息,并将其提供给高级语

义通道,辅助高级语义通道提取图像上下文信息.

考虑到深度可分离卷积(DSC)网络推理速度的优势,本文设计一种基于DSC网络结构的辅助语义模块(Aux);然后,将3个(Aux)模块进行串联,并在尾部添加一种全连接结构的注意力机制模块(Atten),以此保留最大的感受野,输出更完整的辅助特征信息;最后,将各阶段提取的辅助语义特征逐层传递给高级语义模块,构建辅助通道,实现不同语义信息间桥梁的作用,网络结构如图5所示.

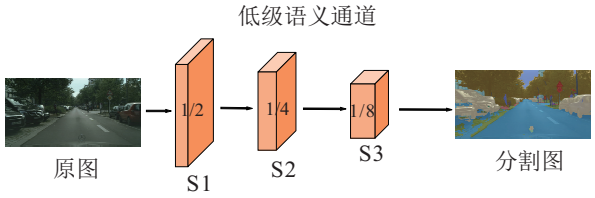


图5　辅助语义通道网络结构

图5中Aux模块由3层DSC结果和1层步长为2的3×3卷积结果结合后输出,Atten模块先将输入变为$n \times 1\,000 \times 1 \times 1$($n$为类别数)的矩阵结构,再对其进行步长为1的1×1卷积,最后再恢复成原来的输入形状并与输入进行合并后输出.

设输入辅助语义通道的特征矩阵为$K_{m\times n}$,Aux模块用$\mathrm{Aux}(x)$函数表示,Atten模块用$\mathrm{Atten}(x)$表示,用$\mathrm{AU}(x)$表示辅助语义通道信息提取过程,则辅助语义通道信息提取过程可表达为

$$\mathrm{AU}(K_{m\times n}) = \mathrm{Atten}(\mathrm{Aux}(\mathrm{Aux}(\mathrm{Aux}(K_{m\times n})))). \quad (4)$$

由图5可知,辅助语义通道最终输出$H/16 \times W/16 \times 64$的过渡语义特征$\mathrm{AU}(K_{m\times n})$.

进一步,设上采样操作为$\mathrm{Up}(x,k)$(k为上采样倍数,此处$k=4$),辅助语义通道各阶段输出的辅助语义信息可表达为

$$\mathrm{Aux1}(K_{m\times n}) = \mathrm{Aux}(K_{m\times n}), \quad (5)$$

$$\mathrm{Aux2}(K_{m\times n}) = \mathrm{Aux}(\mathrm{Aux1}(K_{m\times n})), \quad (6)$$

$$\mathrm{Aux3}(K_{m\times n}) = \mathrm{Aux}(\mathrm{Aux2}(K_{m\times n})), \quad (7)$$

$$\mathrm{Aux1}'(K_{m\times n}) = \mathrm{Up}(\mathrm{AU}(K_{m\times n}),4). \quad (8)$$

2.2.3　高级语义通道

高级语义通道主要提取图像中上下文逻辑关系及完整类别语义信息. 先设计启动模块(Start)和特征聚合模块(FGB);再将启动模块、特征聚合模块以及接收自辅助语义通道的反馈信息进行三者融合,实现对图像深层语义信息的提取;最后利用语义嵌套模块(SEB)将辅助语义通道与高级语义通道的深层语义信息进行整合,完成高级语义通道的信息输出过程. 网络结构如图6所示.

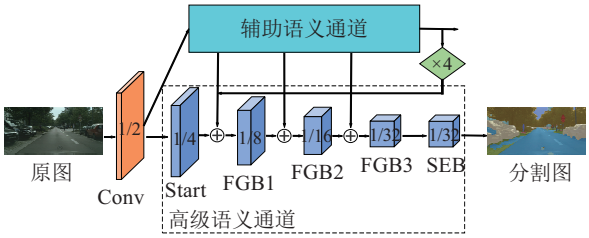


图6　高级语义通道网络结构

1) 启动模块(Start):该通道启动模块为第1阶段,如图7所示,使用了卷积和最大池化两种下采样方式,将输出结果连接再卷积输出,以此增强特征表达能力. 设函数$\mathrm{ST}(x)$为启动模块的推理过程.

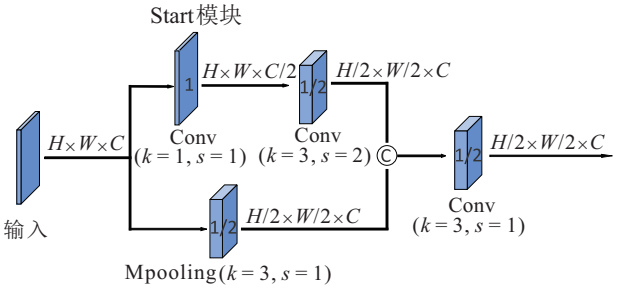


图7　Start模块网络结构

2) 特征聚合模块(FGB):在启动模块之后是特征聚合模块,特征聚合模块由两个特征聚合子模块组成,特征聚合子模块网络结构如图8所示. 子模块也采用深度卷积结构,增加特征层数,提取更深层语义信息,以便更有效地将语义特征聚合并输出.

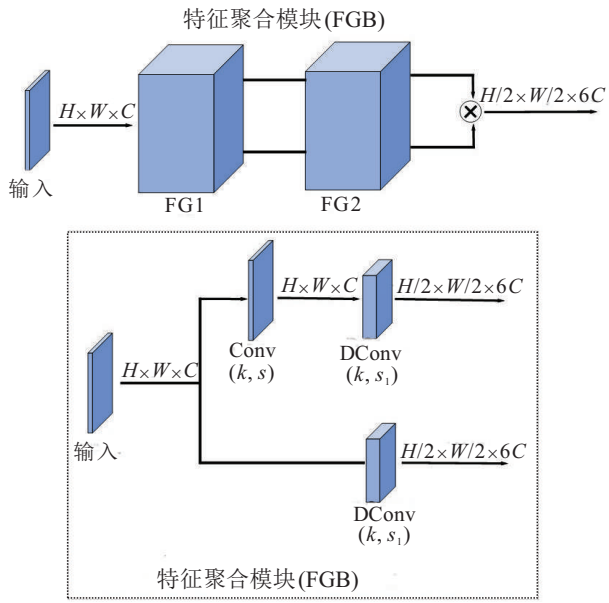


图8　特征聚合模块网络结构

假设$\mathrm{FG}(x,s_1)$表示特征聚合子模块推理过程,$\mathrm{FGB}(x)$表示特征聚合模块推理过程,则聚合过程表达为

$$\mathrm{FGB}(K'_{m\times n}, s_1, s'_1) = \mathrm{FG}(\mathrm{FG}(K'_{m\times n}, s_1), s'_1). \quad (9)$$

其中: $K'_{m\times n}$ 为输入特征矩阵,步长 $s_1 = 2, s'_1 = 1$.

进一步结合式(5)～(9),可将FGB1、FGB2、FGB3推理过程表达如下:

$$\mathrm{FGB1}(K_{m\times n}) = \mathrm{FGB}(\mathrm{ST}(K_{m\times n}) + \mathrm{Aux1}(K_{m\times n}) + \mathrm{Aux}'1(K_{m\times n}), s_1, s'_1), \quad (10)$$

$$\mathrm{FGB2}(K_{m\times n}) = \mathrm{FGB}(\mathrm{FGB1}(K_{m\times n}) + \mathrm{Aux2}(K_{m\times n}), s_1, s'_1), \quad (11)$$

$$\mathrm{FGB3}(K_{m\times n}) = \mathrm{FGB}(\mathrm{FGB2}(K_{m\times n}) + \mathrm{Aux3}(K_{m\times n}), s_1, s'_1). \quad (12)$$

3) 语义嵌套模块(SEB): 该通道最后为一个语义嵌套模块,该模块使用全局平均池化和跳跃连接结构,将高级语义通道和辅助语义通道信息进行深度融合,从而更有效地嵌入全局上下文逻辑关系,如图9所示.

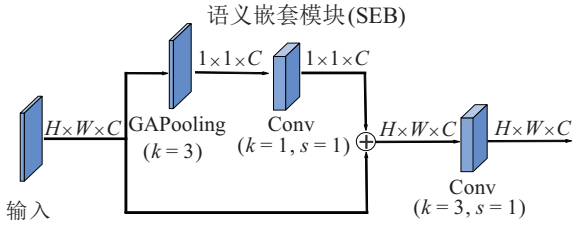


图 9 语义嵌套模块网络结构

设 $\mathrm{SEB}(x)$ 表示该模块的推理过程, $\mathrm{AS}(x)$ 为高级语义通道语义信息提取过程,则结合式(12)可将 $\mathrm{AS}(x)$ 推理过程表达如下:

$$\mathrm{AS}(K_{m\times n}) = \mathrm{SEB}(\mathrm{FGB3}(K_{m\times n})). \quad (13)$$

综上,可由高级语义通道提取 $H/32 \times W/32 \times 128$ 的高级语义特征 $\mathrm{AS}(K_{m\times n})$.

2.3 3类语义特征加权聚合

上述3类语义信息本质上均表征的是图像的局部语义特征,且具有互补性,为此需要将三者进行融合得到图像的全局语义描述. 为此,本节提出一种特征加权聚合方法实现3类语义快速融合. 具体融合原理如图10所示.

1) 加权聚合原理分析.

设3类语义特征加权权重分别为 ε_1、ε_2 和 ε_3,其初值均设为1. 运行中权重的改变由3类语义特征通过验证集中测得的 $\mathrm{MIoU}^{[21]} = \{\mathrm{MIoU1}, \mathrm{MIoU2}, \mathrm{MIoU3}\}$ 自适应决定. 自适应更新规则如下: 采用当前各通道的网络权重在验证集上求得MIoU,再

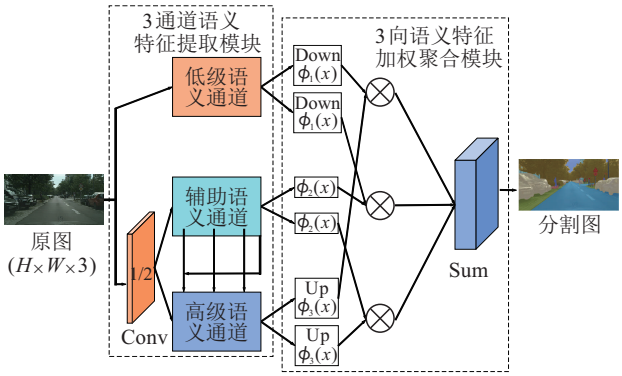


图 10 3类语义特征加权聚合模块

由MIoU值根据下式进行更新:

$$\varepsilon_i = \begin{cases} 1 + \mathrm{MIoU}, & \mathrm{MIoU}i = \max\{\mathrm{MIoU}\}; \\ 1, & \min\{\mathrm{MIoU}\} < \mathrm{MIoU}i < \max\{\mathrm{MIoU}\}; \\ 1 - \mathrm{MIoU}, & \mathrm{MIoU}i = \min\{\mathrm{MIoU}\}. \end{cases} \quad (14)$$

其中 $i = 1, 2, 3$. 当某个通道的MIoU值为3个通道的最大值时,表明该通道特征提取能力较强,用 $(1 + \mathrm{MIoU})$ 作为权重以增强该通道所提取的特征信息;反之,当某个通道的MIoU值为3个通道的最小值时,表明该通道特征提取能力较弱,用 $(1 - \mathrm{MIoU})$ 作为权重以减弱该通道所提取的特征信息;除这两种情况,权重仍保持为上一次的值.

图10中Down是一个步长为2的3×3卷积, Up是一个放大倍数为2的双线性插值上采样,设 $\mathrm{Up}(x)$ 表示上采样过程,3类语义特征加权过程为

$$\begin{cases} \phi_1(x) = \varepsilon_1 \times \mathrm{Conv}(x), \\ \phi_2(x) = \varepsilon_2 \times \mathrm{sigmoid}(x), \\ \phi_3(x) = \varepsilon_3 \times \mathrm{Up}(x). \end{cases} \quad (15)$$

其中

$$\mathrm{sigmoid}(x) = \frac{1}{1 + \mathrm{e}^{-x}}. \quad (16)$$

2) 3类语义特征聚合.

图10中⊗表示将两种结果相乘后再进行步长为1的3×3卷积,卷积用 $\mathrm{Conv1}(x)$ 表示,3类语义特征两两聚合过程用 $\mathrm{SA1}(x)$、$\mathrm{SA2}(x)$ 和 $\mathrm{SA3}(x)$ 表示,复合过程用 $\mathrm{MSA}(x)$ 表示,由式(15)可得

$$\begin{cases} \mathrm{SA1}(x) = \mathrm{Conv1}(\phi_1(x) \times \phi_3(x)), \\ \mathrm{SA2}(x) = \mathrm{Conv1}(\phi_1(x) \times \phi_2(x)), \\ \mathrm{SA3}(x) = \mathrm{Conv1}(\phi_2(x) \times \phi_3(x)). \end{cases} \quad (17)$$

3类语义特征最后的聚合过程可表述如下:

$$\mathrm{MSA}(x) = \mathrm{SA1}(x) + \mathrm{SA2}(x) + \mathrm{SA3}(x). \quad (18)$$

综上,式(15)～(18)完成了加权聚合的推理.

2.4　增强训练机制模块

为进一步提高训练速度和训练效果,本文还设计一种新的增强训练机制模块,可强化训练阶段特征表示. 如图11所示,在训练MCDWA_Net网络时,将各通道输出特征进行上采样与网络最终输出结果共同计算损失函数,并进行随机梯度下降学习.

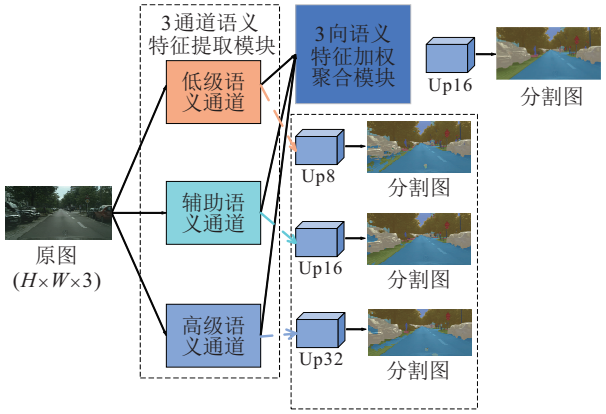


图 11　增强训练模块

基于深度学习的语义分割算法中,一般采用Cross-Entropy或Diceloss[22]损失函数进行梯度下降学习,但Cross-Entropy在训练语义分割模型时,权重更新会受到误差的影响,如图像中前景像素远小于背景像素,则会使训练出的模型偏向背景,即Cross-Entropy相对于背景较小的分割目标训练效果较差,Dice loss在训练语义分割模型时会对反向传播不利,使训练过程不稳定,但其能够克服图像中前景像素与背景像素数量不均衡的影响. 因此,为克服上述两种损失函数的缺陷,本文设计一种像素信息熵损失函数(pixel information entropy),简称PIE loss函数,通过对比图像整体像素相似性和各类别像素相似性关系来计算预测图像和真实图像之间的差距,并将这种差距作为最终的损失函数. PIE loss函数定义如下.

首先,设$x = x_1, x_2, \ldots, x_n$为预测数据,$y = y_1, y_2, \ldots, y_n$为真实数据,$n$为类别数,则PIE loss指标计算公式定义如下:

$$\mathrm{PIE}(x,y) = \alpha\left(1 - \frac{2|x \bigcap y|}{|x| + |y|}\right) - \sum_{i-1}^{n} y_i \log(x_i). \quad (19)$$

其中: $1 - \dfrac{2|x \bigcap y|}{|x| + |y|}$表征真实数据与预测数据的总体像素相似性; $-\sum_{i-1}^{n} y_i \log(x_i)$代表真实数据与预测数据之间各类别像素的相似性,二者的有机融合在一定程度上可以兼顾全局与细节特征信息.

式(19)中α为权值,α的计算公式为

$$\alpha = \frac{n-k}{n+k}, \quad (20)$$

其中k为超参数($-n < k \leqslant n$且$k \in z$). 在训练过程中α根据实际情况而定,具体规则如下.

1) 假设语义分割中分割类别数为$n = 2$,此时相当于一个二分类问题,则只需要考虑两个类别,此时真实数据与预测数据之间各类别的像素差距越小越好,因此$-\sum_{i-1}^{n} y_i \log(x_i) - \sum_{i-1}^{n} y_i \log(x_i)$部分占主导地位. 当$\alpha$值越小,$-\sum_{i-1}^{n} y_i \log(x_i)$部分越突出,因此,$k$越大越好,可选$k = 2$,则$\alpha = 0$.

2) 多分类问题语义分割n大于2,由于类别较多,仅依靠像素类别差距无法更好地计算真实数据与预测数据之间的误差,因此,从整体上考虑真实数据与预测数据之间的误差. 为此,设计了$1 - \dfrac{2|x \bigcap y|}{|x| + |y|}$部分进行计算,$1 - \dfrac{2|x \bigcap y|}{|x| + |y|}$部分能够更好地表达真实数据与预测数据之间的全局差距. 此时α值越大越突出$1 - \dfrac{2|x \bigcap y|}{|x| + |y|}$部分,选择$k$越小越好.

3) 当语义分割类别$n = 2$时,选择$k = 2$,则$\alpha = 0$;当$n > 2$时,n增大而k减小,使α不断增大,在实际应用中,α需根据实际情况设定上限.

最后,将各通道输出特征进行上采样,并与网络输出结果共同计算损失,PIE loss表达式为

$$\mathrm{loss}(P_re, A_1, A_2, A_3, T) = \mathrm{PIE}(P_re, T) + \sum_{i=1}^{3} \mathrm{PIE}(A_i, T). \quad (21)$$

其中: T为真实标签,P_re为MCDWA_Net网络输出标签,A_1、A_2和A_3分别为低级、辅助和高级语义通道的输出标签.

3　实验分析与验证

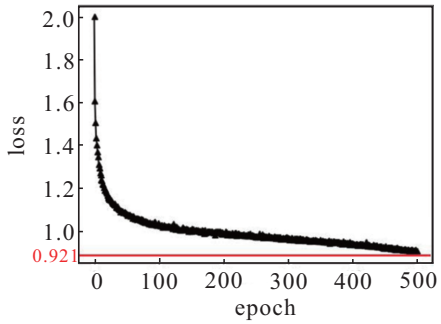
3.1　Cityscapes街景数据集实验

3.1.1　模型训练

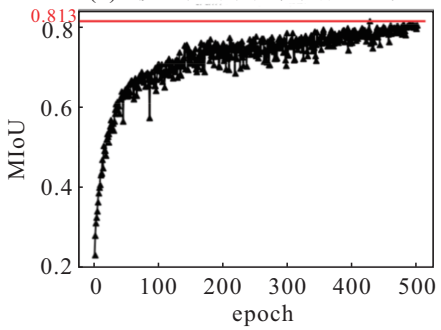
在用Cityscapes[23]街景数据集训练MCDWA_Net模型时,训练批量Batch_size为4,类别数Num_classes为19,迭代次数为Epoches 500,学习策略采用随机梯度下降(SGD)算法,设其动量为0.9,初始学习率为0.05,权重衰减率为0.000 1.

在训练模型时,应用式(21)计算损失函数,并采用随机梯度下降法进行训练学习,训练过程中损失值

变化如图12(a)所示.



(a) 模型训练损失值变化曲线



(b) 模型训练MIoU 值变化曲线

图 12 模型训练值变化曲线

3.1.2 模型性能评价指标

首先引入平均交并比(mean intersection over union, MIoU)对模型的准确性进行评估, MIoU评估的是模型分割物体的精度, MIoU值越高表示物体分割效果越好. 计算方法如下:

$$\mathrm{MIoU}=\frac{1}{K+1}\sum_{i=0}^{k}\frac{\mathrm{TT}_i}{\mathrm{FT}_i+\mathrm{FF}_i+\mathrm{TT}_i}. \tag{22}$$

其中: k为类别数, FT_i表示第i个类别中预测错误且预测为真的样本数, FF_i表示第i个类别中预测错误且预测为假的样本数, TT_i表示第i个类别中预测正确的样本数.

本文算法在Cityscapes街景数据集上训练过程MIoU值变化如图12(b)所示.

3.1.3 模型消融实验

在Cityscapes街景数据集上完成消融实验, 验证网络模型中各模块的有效性, 并利用Cityscapes街景验证集进行算法评估.

表2中前3行表示仅使用一个通道时的分割精度和计算量(GFLOPs). 低级语义通道单独运行时, MIoU值较低, 计算量较大; 辅助语义通道单独运行时, 仅能提取图像的过渡语义信息. 运算过程中各层生成的特征图分辨率较低, 信息量较少, 使计算量偏少, MIoU值偏低. 高级语义通道单独运行时, 计算量一般, 但MIoU值是3通道中最大的, 主要因为高级语义通道包含图像的上下文逻辑关系和较为完整的类别语义特征, 但其缺乏边缘、颜色、结构等低级语义信息, 获取信息也并不完整. 显然, 3个语义通道各有所长, 且都无法获取图像完整的语义信息, 为此可以将3通道有机融合, 从而获得更高分割精度. 此外, 融合方法也具有至关重要的作用, 如表2第4行的直接相加融合效果比第5行的语义特征加权融合效果差很多, 其MIoU值偏低. 但总体而言, 融合后分割效果均优于各通道的单独分割效果.

表 2 MCDWA_Net在Cityscapes街景数据集的消融实验

通道类型			融合方法		训练策略		性能指标	
低级语义	辅助语义	高级语义	相加聚合	3类语义特征加权聚合	普通训练	增强训练	MIoU/%	GFLOPs
√							60.3	15.26
	√						56.8	1.6
		√					66.2	8.25
	√	√	√				76.1	22.1
	√	√		√			78.2	22.4
√	√	√		√	√		79.8	22.4
√	√	√		√		√	80.4	22.4

另外, 模型训练方法对分割效果也有一定影响, 实验中, 采用普通训练方法获得的模型MIoU值比本文提出的增强训练机制获得的MIoU值低0.6%. 一定程度上表明本文所设计的增强训练机制也能提高模型的分割精度.

以图13场景进行消融实验结果为例进一步说明算法的优势. 首先, 如图13所示, 当3个通道单独进行语义分割时, 低级语义通道可以更好地分割出红圈1中的“路灯”, 而另外两个通道中对“路灯”的分割效果较差. 其次, 如图13中的红圈2所示, 高级语义通道可以更好地提取“车”和“地面”的语义信息, 能够更好地提取图像的大物体的类别语义信息. 最后, 辅助语义通道在一些大物体或者小物体的一些部位可以提取到高级和低级语义通道都无法提取的语义特征信息, 如图13中的红圈3所示, 辅助语义通道可提取到低级语义通道无法提取的小物体“人”的语义信

息,也可以提取到高级语义通道无法提取到的大物体“车”的部分语义信息. 由图13中的多通道深度加权聚合网络得到的语义分割图显然可见红圈1、2、3中的分割效果都较好,能明显看出3类语义特征加权聚合模块将3通道提取的3类语义信息加权聚合到一起,从而得到更加完整的全局语义信息.

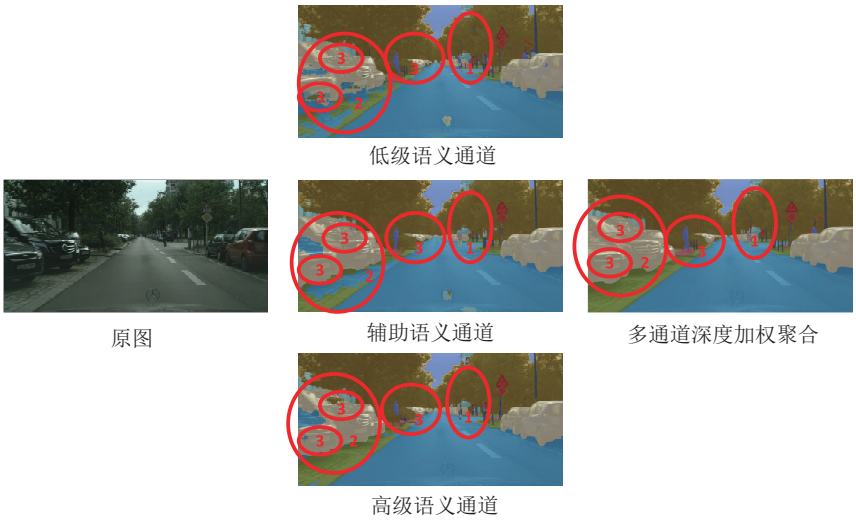


图13　消融实验效果图

3.1.4　模型性能对比实验

将BiSeNet[14]、BiSeNetV2[19]、DFANet[15]、Deeplab V3[24]、ShuffleNet V2[25]和SAB Net[26]算法利用相同设备在Cityscapes街景数据集上训练出相应模型,其训练相关参数设置如3.1.1节所述,采用传统训练策略进行训练. 表3展现了7种算法性能指标对比结果. 由表3可知,本文的MCDWA_Net算法分割精度均明显高于其他5种算法,其运算复杂度与BiSeNet、BiSeNetV2、DFANet算法相近,推理速度稍稍滞后于这3种算法;相比于Deeplab V3、ShuffleNet V2和SAB Net算法,在算法精度和推理速度上均优于这3种算法,虽然本文算法在精度上略低于SAB Net算法,但在速度上大幅超过SAB Net算法.

表3　不同算法在Cityscapes街景数据集上的性能比较

算法名称	MIoU/%	GFLOPs	速度/(ms/帧)
MCDWA_Net	80.4	22.4	16
BiSeNet[14]	68.1	14.8	13
BiSeNetV2[19]	72.6	21.15	15
DFANet[20]	71.1	3.4	12
Deeplab V3[24]	71.4	185.2	157
ShuffleNetV2[25]	70.3	128.71	64.9
SAB Net[26]	82	156.35	101

7种算法分割效果如图14所示,分割效果差别如图中红圈所示. 由表3和图14可知,本文所提出算法在保证实时推理速度的前提下,大幅提升了分割精度,其MIoU值可达80.4%,推理速度为16 ms/帧,满足实时性要求. 因此,本文算法综合性能更优、性价比更高,具有更好的实际应用价值.

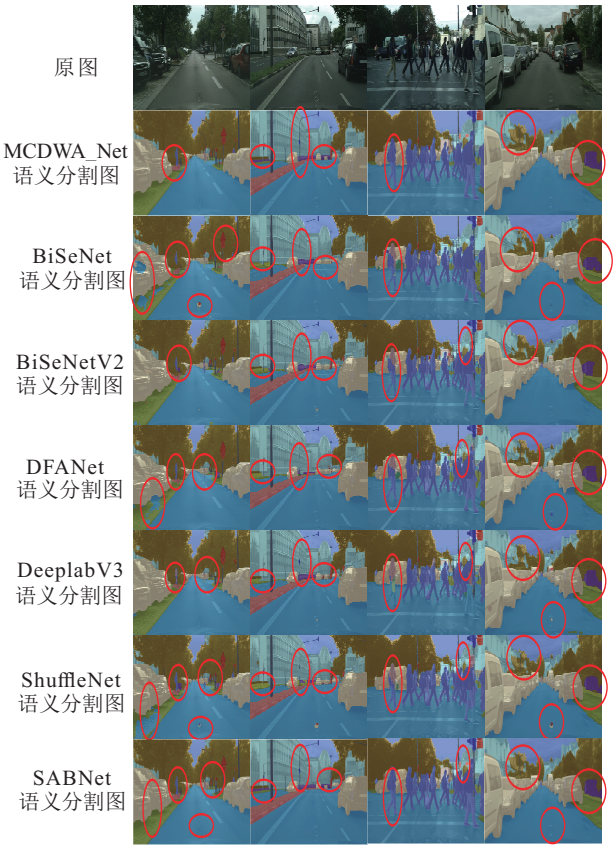


图14　7种语义分割算法实际分割效果图

3.2　实验室走廊场景数据集实验

实验室走廊场景数据集是自建场景数据集,目的是验证本文设计的增强训练模块的训练效果. 采用走廊场景数据集训练MCDWA_Net模型时,设置训练批量Batch_size为4,类别数Num_classes为8,其余参

数均与3.1.1节相同.

采用本文所提出的增强训练PIE loss函数与单独使用Cross Entropy和Dice loss函数进行对比实验, 验证PIE loss的有效性. 训练中损失值变化如图15所示, MIoU值变化如图16所示, 各种损失函数的训练效果对比结果如表4所示.

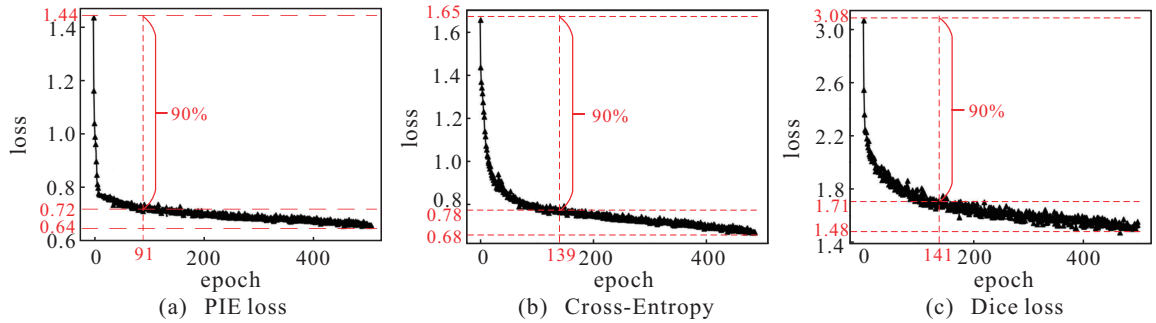


图 15 模型训练损失值loss变化曲线

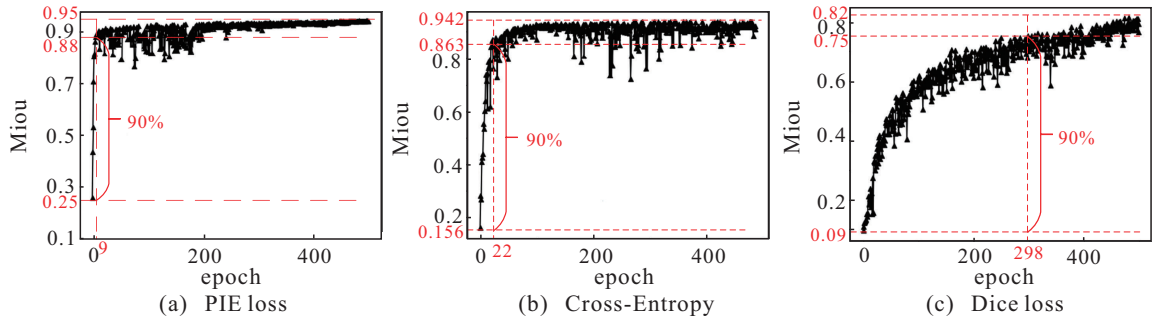


图 16 模型训练MIoU值变化曲线

表 4 3种损失函数训练效果对比结果

损失函数	loss收敛90%/次	MIoU/%
PIE loss	90	95
Cross-Entropy	139	91.2
Dice loss	141	82

备注: loss收敛90次(最少次数)

结合图15和表4可知, 本文的PIE loss在训练过程中循环训练90次即可让损失值降低90%以上, 而Cross Entropy在训练过程中需要循环139次才能使损失值降低90%, Dice loss在训练过程中需要循环141次才能使损失值降低90%. 由此可见, 本文所设计损失函数能够更快降低训练时的损失值.

结合图16和表4可知, 在走廊场景数据集训练过程中, 使用PIE loss能够使模型MIoU达到95%, 且仅需要循环训练9次即可使模型的MIoU提升90%, 而ross Entropy与Dice loss能够使模型精度最高仅为91.2%和82%, 且分别需要循环训练28次和298次才能使模型的MIoU提升90%. 由此可见, 本文所设计的强化训练机制能够更快地使语义分割网络模型提升至更高精度, PIE loss优势明显.

本文也将BiSeNet[14]、BiSeNetV2[19]、DFANet[15]、Deeplab V3[24]、ShuffleNet V2[25]和SAB Net[26]算法利用相同设备在实验室走廊场景数据集上训练出相应模型, 并与本文算法性能进行比较, 7种算法在实验室走廊场景数据集上的性能如表5所示, 其对实验室走廊场景的分割效果如图17所示.

表 5 不同算法在走廊场景数据集性能比较

算法名称	MIoU/%	GFLOPs	速度/(ms/帧)
MCDWA_Net	93.2	22.4	16
BiSeNet[14]	79.9	14.8	13
BiSeNetV2[19]	82.3	21.15	15
DFANet[20]	81.4	3.4	10
Deeplab V3[24]	87.6	185.2	167
ShuffleNetV2[25]	81.9	128.71	64.9
SAB Net[26]	94.5	156.35	101

由表5可知, 本文所提出的MCDWA_Net算法分割精度仅次于SAB Net算法, 相差不大, 但其推理速度和计算量都显著小于SAB Net算法; 其推理速度与BiSeNet、BiSeNetV2和DFANet算法非常接近, 略慢于这3种算法, 综合比较则本文算法优势明显.

由图17可知(左图为较简单场景, 右图为较复杂场景), 在左图中本文算法的优越性并不十分显著, 但在右图中红圈标注处, 本文算法与SAB Net算法明显优于其他算法, 其他5种算法对场景的语义分割均有一定的缺陷. 因此, 针对实验室走廊场景的实验中, 本文所提算法的综合性能更优, 间接表明本文算法在实

际场景中具有一定的优越性.

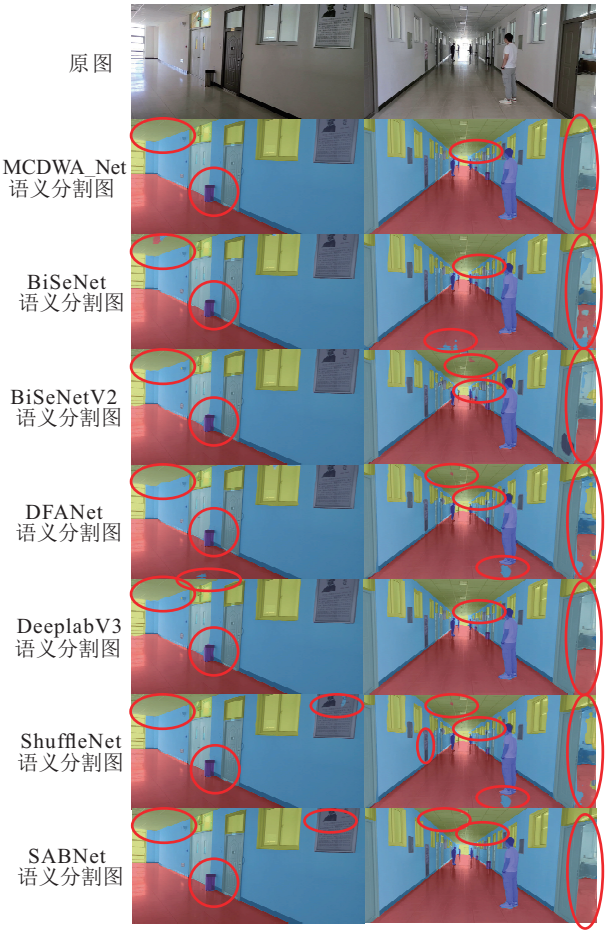


图 17　7种语义分割算法分割效果图

4　结　论

本文提出的MCDWA_Net,具有如下特点:

1)通过3通道结构分别提取图像的3类互补语义信息.

2)利用3类语义特征加权聚合方法,将3类互补语义信息加权聚合,从而得到更加全面的图像语义特征.

3)引入一种增强训练方法,使MCDWA_Net算法在训练过程中获得更好的训练效果.

算法在公共数据集Cityscapes街景数据集和自建实验室走廊场景数据集上进行了验证,结果表明MCDWA_Net算法具有较快的推理速度,可达16 ms/帧,且在Cityscapes数据集中具有高达80.4%的识别精度,在走廊场景数据集达到93.2%的识别精度. 因此,本算法能够很好地实现图像语义分割速度和分割精度之间的兼顾,具有较好的推广价值.

参考文献(References)

[1] 黄庭鸿, 聂卓赟, 王庆国, 等. 基于区块自适应特征融合的图像实时语义分割[J]. 自动化学报, 2021, 47(5): 1137-1148.
(Huang T H, Nie Z Y, Wang Q G, et al. Real-time image semantic segmentation based on block adaptive feature fusion[J]. Acta Automatica Sinica, 2021, 47(5): 1137-1148.)

[2] Zhou B L, Zhao H, Puig X, et al. Semantic understanding of scenes through the ADE20K dataset[J]. International Journal of Computer Vision, 2019, 127(3): 302-321.

[3] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, 2012: 3354-3361.

[4] 余烨, 傅云翔, 杨昌东, 等. 基于FR-ResNet的车辆型号精细识别研究[J]. 自动化学报, 2021, 47(5): 1125-1136.
(Yu Y, Fu Y X, Yang C D, et al. Fine-grained car model recognition based on FR-ResNet[J]. Acta Automatica Sinica, 2021, 47(5): 1125-1136.)

[5] Wu Z F, Shen C H, Hengel A V D. Real-time semantic image segmentation via spatial sparsity[J/OL]. 2017, arXiv: 1712.00213.

[6] Paszke A, Chaurasia A, Kim S, et al. ENet: A deep neural network architecture for real-time semantic segmentation[J/OL]. 2016, arXiv: 1606.02147.

[7] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.

[8] 韩冲, 王俊丽, 吴雨茜, 等. 基于神经进化的深度学习模型研究综述[J]. 电子学报, 2021, 49(2): 372-379.
(Han C, Wang J L, Wu Y X, et al. A review of deep learning models based on neuroevolution[J]. Acta Electronica Sinica, 2021, 49(2): 372-379.)

[9] Evans B, Al-Sahaf H, Xue B, et al. Evolutionary deep learning: A genetic programming approach to image classification[C]. 2018 IEEE Congress on Evolutionary Computation. Rio de Janeiro, 2018: 1-6.

[10] Wang C Y, Xu C, Yao X, et al. Evolutionary generative adversarial networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(6): 921-934.

[11] Assuncao F, Sereno D, Lourenco N, et al. Automatic evolution of AutoEncoders for compressed representations[C]. 2018 IEEE Congress on Evolutionary Computation. Rio de Janeiro, 2018: 1-8.

[12] Kozma R, Alippi C, Choe Y, et al. Artificial intelligence in the age of neural networks and brain computing[M]. Academic Press, 2018: 293-312.

[13] Gaier A, Ha D. Weight agnostic neural networks[J/OL]. 2019, arXiv:1906.04358.

[14] Xu Q, Ma Y N, Wu J, et al. Faster BiSeNet: A faster

bilateral segmentation network for real-time semantic segmentation[C]. 2021 International Joint Conference on Neural Networks. Shenzhen, 2021: 1-8.

[15] Li H C, Xiong P F, Fan H Q, et al. DFANet: Deep feature aggregation for real-time semantic segmentation[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 9514-9523.

[16] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[C]. IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE, 2016: 640-651.

[17] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[18] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[19] Yu C Q, Gao C X, Wang J B, et al. BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vision, 2021, 129(11): 3051-3068.

[20] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 1800-1807.

[21] Garcia-Garcia A, Orts-Escolano S, Oprea S, et al. A review on deep learning techniques applied to semantic segmentation[J/OL]. 2017, arXiv: 1704.06857.

[22] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]. 2016 Fourth International Conference on 3D Vision (3DV). Stanford, 2016: 565-571.

[23] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 3213-3223.

[24] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[25] Türkmen S, Heikkilä J. An efficient solution for semantic segmentation: ShuffleNet V2 with atrous separable convolutions[C]. Scandinavian Conference on Image Analysis. Cham: Springer, 2019: 41-53.

[26] Ding X F, Shen C M, Zeng T Y, et al. SAB net: A semantic attention boosting framework for semantic segmentation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, PP(99): 1-13.

作者简介

齐咏生 (1975−), 男, 教授, 博士, 从事机器人协同控制技术、机器人视觉算法等研究, E-mail: qys@imut.edu.cn;

陈培亮 (1997−), 男, 硕士, 从事机器人视觉算法的研究, E-mail:cpl2gyf@163.com;

高学金 (1973−), 男, 教授, 博士, 从事发酵过程建模、监控与故障诊断等研究, E-mail: gaoxuejin@bjut.edu.cn;

董朝轶 (1976−), 男, 教授, 博士, 从事地面移动机器人的动态建模、导航等研究, E-mail: 16136745@qq.com;

魏淑娟 (1998−), 女, 硕士生, 从事机器视觉的研究, E-mail: shujuanwei_1022@163.com.